https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

Low-Cost Air Quality Sensor Evaluation and Calibration in Contrasting Aerosol Environments

Pawan Gupta[1,2], Prakash Doraiswamy[3], Jashwanth Reddy[4,1], Palak Balyan[5,6], Sagnik Dey[5], Ryan Chartier[3], Adeel Khan[7], Karmann Riter[3], Brandon Feenstra[8], Robert C. Levy[9], Nhu Nguyen Minh Tran[4], Olga Pikelnaya[8], Kurinji Selvaraj[8], Tanushree Ganguly[7], Karthik Ganesan[7]

[1]STI-Universities Space Research Associations, Huntsville, AL, USA.
[2] NASA Marshall Space Flight Center, Huntsville, AL, USA
[3]RTI International, Research Triangle Park, NC, USA.
[4]The University of Alabama in Huntsville, AL, USA.
[5]Indian Institute of Technology, New Delhi, India
[6]Health Effects Institute (HEI), Boston, U.S.A.
[7]Council on Energy, Environment, and Water (CEEW), New Delhi, India
[8]South Coast Air Quality Management District, CA, USA.
[9]NASA Goddard Space Flight Center, Huntsville, AL, USA.

Correspondence to: Pawan Gupta (pawan.gupta@nasa.gov)

Abstract. The use of low-cost sensors (LCS) in air quality monitoring has been gaining interest across all walks of society, including community and citizen scientists, academic research groups, environmental agencies, and the private sector. Traditional air monitoring, performed by regulatory agencies, involves expensive regulatory-grade equipment and requires ongoing maintenance and quality control checks. The low-price tag, minimal operating cost, ease of use, and open data access are the primary driving factors behind the popularity of LCS. This study discusses the role and associated challenges of $PM_{2.5}$ sensors in monitoring air quality. We present the results of evaluations of the PurpleAir (PA.) PA-II LCS against regulatory-grade $PM_{2.5}$ federal equivalent methods (FEM) and the development of sensor calibration algorithms. The LCS calibration was performed for 2 to 4 weeks during December 2019-January 2020 in Raleigh, NC, and Delhi, India, to evaluate the data quality under different aerosols loadings and environmental conditions. This exercise aims to develop a robust calibration model that uses PA measured parameters (i.e., $PM_{2.5}$, temperature, relative humidity) as input and provides bias-corrected $PM_{2.5}$ output at an hourly scale. Thus, the calibration model relies on simultaneous measurements of $PM_{2.5}$ by FEM as target output during the calibration model development process. We applied various statistical and machine learning methods to achieve a regional calibration model. The results from our study indicate that, with proper calibration, we can achieve bias-corrected $PM_{2.5}$ data using PA sensors within 12% percentage mean absolute bias at hourly and within 6% for a daily average. Our study also suggests that pre-deployment calibrations developed at local or regional scales should be performed for the PA sensors to correct data from the field for scientific data analysis.

1. Introduction

Air quality monitoring is critical for managing and mitigating air pollution at varying spatiotemporal scales. However, air quality monitoring is limited in many parts of the world (Martin et al., 2019) in part due to the high cost and technical experience requirements of operating regulatory-grade monitors (R.G.M.). Regulatory-grade continuous air quality monitors have high measurement accuracy under varying operating conditions. The high cost of RGM and their associated infrastructure needs and regular maintenance also limit the extensive deployment of such monitors in a region and the spatial density of the network. This is particularly true in developing countries. The lack of data affects critical decision-making by the public about their day-to-day activities and regulatory agencies for controlling and mitigating air pollution in many regions.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

In recent years, low-cost sensors (LCS) are increasingly being used for monitoring pollution at the local scale and have been suggested as one component of a hybrid monitoring system (Martin et al., 2019). LCS that report fine particulate matter ($PM_{2.5}$, i.e., particles with aerodynamic diameter less than or equal to 2.5 micrometers) in particular have shown increasing the potential for air quality monitoring along with RGM measurements and satellite observations (Feenstra et al., 2019; Gupta et al., 2018; Wallace et al., 2021). LCS technologies have significantly lower costs (~100 times cheaper equipment cost), minimal infrastructure requirements (low power requirements and wireless internet/S.D. card), and smaller footprints. They are easy to install, and in some cases, data are uploaded to cloud services in real-time. The relatively lower cost also allows the deployment of LCS at multiple locations to fill gaps in ground monitoring and better assess the spatial variability of air pollutants at higher time resolutions. In addition to LCS, the hybrid approach includes satellite observations. The Satellite-based measurements offer global spatial coverage that may help fill in spatial gaps in an air monitoring network (van Donkelaar et al., 2015; van Donkelaar et al., 2016). However, most satellite data are limited in temporal coverage (e.g., once a day except for geostationary satellites), impacted by cloud cover, and potentially prone to larger uncertainties in regions with limited or no ground-based monitors.

While several LCS are commercially available for indoor and outdoor air pollution measurements, their performance varies compared to RGM. The South Coast Air Quality Management District (South Coast AQMD) Air Quality Sensor Evaluation Center (AQ-SPEC) evaluates the performance of air quality sensors in a laboratory under controlled conditions and in ambient environments and provides the evaluation reports on its website (South Coast AQMD, 2022). In addition to LCS intrinsic limitations, the particle sources, types, concentration, seasonality, and weather conditions impact the measurement uncertainty (Wang et al., 2015). Over the past few years, many published studies have evaluated and developed correction factors for various LCS with encouraging results (Badura et al., 2019; Bi et al., 2020a; Bi et al., 2020b; Di Antonio et al., 2018; Si et al., 2020; Wallace et al., 2021; Wang et al., 2020; Zusman et al., 2020). Most studies concluded that after calibration, data from LCS showed better agreement with reference methods and reduced measurement error, making them suitable for various applications. For example, Bi et. al., (2020b) found that integrating data from LCM and RGM as part of a geographically weighted regression model improved the spatial representation of $PM_{2.5}$ predictions and identifying hotspots such as wildfires. Wang et al. (2020) applied the correction algorithm to the sensors in their monitoring network for use in citizen science, public education, environmental research, and support policy. Because of their potential, funding agencies around the world are funding various efforts to utilize LCS along with citizen science.

In our NASA-funded citizen science study, we evaluated and deployed LCS as part of a community volunteer-based ambient sensor network to generate spatially and temporally resolved air quality data to refine satellite-based surface $PM_{2.5}$ estimates. The secondary purpose of LCS data in this project is to evaluate spatiotemporal gradients near-surface and in the atmospheric column. Citizen scientists deployed and hosted the sensors. To ensure appropriate data quality and assess spatial gradients, it was important to calibrate all sensors on a uniform basis prior to deployment by citizen scientists. In this article, we present our sensor evaluation effort. For our study, we selected the PurpleAir (PA.) PA-II sensor due to its documented performance (Feenstra et al., 2019; South Coast AQMD, 2022). Additionally, PA-II was selected based on its low cost, ease of use by citizen scientists, and open data framework, which provides a real-time map and opens data access through Application Programming Interfaces (APIs). As of March 2022, the PA network (www.purpleair.com) has more than 21,000 units deployed around the world by individuals, community groups, organizations, and public and private institutions to monitor air quality at their homes, office, and public locations to achieve more spatial and temporal coverage of air quality.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022


Atmospheric Measurement Techniques Discussions

Various research groups have also evaluated the PA units and developed calibration coefficients. Wallace et al., (2021) (and references therein) thoroughly reviewed the literature on testing and calibration methods and proposed their own method. They compared 33 PA units located within 500 meters of US E.PA Air Quality System (AQS) stations over a period of 18 months and proposed a method of converting particle number output from the PA-II to mass concentration, thus avoiding the use of PA data streams "CF=1" and "CF=ATM" provided by the manufacturer. The US E.PA developed a regression-based correction equation for the continental US (CONUS) region by evaluating 53 PA sensors collocated against regulatory-grade continuous $PM_{2.5}$ monitors, referred to as Federal Equivalent Methods (FEM), at 39 locations across 16 states (Barkjohn et al., 2021). Their final correction equation reduced the overall error in daily average concentrations by about 62.5% across the US at an average concentration of 9 $\mu g/m^3$. In India, few studies (Prakash et al., 2021; Zheng et al., 2018) have attempted to evaluate the quality of LCS measurements and developed specific correction equations. Zheng et. al. (2018) used a Plantower sensor (PMS 3003) as part of a custom-made sensor package and evaluated its performance against a FEM in Research Triangle Park, NC (mean $PM_{2.5}$ of 10 ± 3 $\mu g/m^3$) and in Kanpur, India (mean $PM_{2.5}$ 36 ±17 and 116 ± 57 $\mu g/m^3$ during monsoon and post-monsoon seasons respectively). They noted a non-linear response beyond ~125 $\mu g/m^3$ and found that following calibration (quadratic model) and correction for RH, the sensor measurements were within 10% of the reference values. Prakash et. al. (2021) evaluated a different low-cost sensor (APT-MAXIMA) at three sites (urban, industrial, and background) over an 8-month period from May 2019 to February 2020 in Delhi. They found the corrected hourly data to correlate well (R^2 > 0.84) with the reference measurements, with slopes ranging from 0.81 (industrial) to 0.99 (urban).

In this study, we developed a machine learning algorithm (MLA) to calibrate PA-II $PM_{2.5}$ measurements by comparing them against RGM. The MLA uses $PM_{2.5}$ from the CF=ATM stream, along with temperature (T) and relative humidity (RH) measurements from the PA-II sensor as inputs and generates hourly averaged calibrated $PM_{2.5}$ mass concentration. MLA models are developed, tested, and validated independently for two geographic regions with distinct aerosol loadings. Our study provides a novel approach in two ways: 1) multiple (50+) PA-II units were collocated (within a few meters) along with FEM instruments at each location, allowing intercomparison, both among the PA-II units and with the FEM instrument; and 2) collocation was performed in two regions with different environmental conditions (Raleigh area, North Carolina, U.S.A.; and Delhi, India – henceforth, referred to as Raleigh and Delhi, respectively) spread over two continents.

2. $PM_{2.5}$ Measurement Instrumentation

2.1. PA-II Low-cost Sensors

The PA-II (Purple Air L.L.C., Draper, UT, U.S.A.) sensor is an optical particle sensor (OPS) that houses two raw OPS (PMS 5003) manufactured by Plantower (Beijing, China). These two PMS 5003 sensors are referred to as Channel A and Channel B. The PMS 5003 OPS is a nephelometer that measures particle loading through light scattering (wavelength~650 nm) (Hagan and Kroll, 2020a). Sampled air intercepts a beam of light. The light scattered by the ensemble of particles in the sampled air is detected by a photodiode at a roughly 90° angle (Kelly et al., 2017), although Hagan and Kroll (2020b) note that there is no focused collection. The amplitude of the scattered light is measured and correlated to particulate matter (PM) concentration (Hagan and Kroll, 2020a). OPS technology has limitations for measuring PM mass concentrations as changes in aerosol size distributions, aerosol optical properties, and particle density can impact the performance of OPS and lead to measurement errors (Hagan and Kroll, 2020a). The PA-II sensor requires Wi-Fi and power (5V) to be operational. The PA-II, once Wi-Fi is configured and registered with PurpleAir, streams live data to a public map (PurpleAir, 2022). PurpleAir supports open data and provides access to collected data. The primary sensor-reported data include PM_1, $PM_{2.5}$, and PM_{10} concentrations with a factory-specified correction factor for ambient measurements (CF=ATM), concentrations with CF=1 factor recommended by

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022


Atmospheric Measurement Techniques Discussions

the manufacturer for use in indoor measurements, T, and RH. We use $PM_{2.5}$ concentrations from the "CF=ATM" data stream along with T & RH.

2.2. Regulatory-Grade Monitors

Regulatory-grade continuous $PM_{2.5}$ FEMs are those that are certified to be equivalent to a Federal Reference Method. Each collocation site used a different FEM. The measurement site at Delhi employed the Met One BAM-1022 (Met One Instruments, Inc., Grants Pass, OR), while the site in Raleigh ran the Teledyne T640x (Teledyne API, San Diego, CA) sampler. A brief description of the FEMs is provided below.

Met One BAM-1022: The Met One BAM (beta-attenuation monitor) measures the attenuation of beta radiation due to PM. Ambient air is sampled through a US E.PA approved $PM_{2.5}$ inlet at 16.7 L/min. The sampled particles are deposited on a filter tape located between a beta radiation source and a detector. The detector measures the change in intensity of beta radiation due to particles deposited on the filter. The BAM-1022 takes a differential measurement between the beginning and end of a sampled time period. The BAM reported data at a 1-hr resolution. The BAM is inspected every week and cleaned if required. Calibration is performed on an annual basis.

Teledyne T640x: The T640x sampler is an optical aerosol spectrometer. A polychromatic light source shines a light on the preconditioned ambient sample. The scattered light intensity is detected at a 90° angle at the single particle level. The T640x determines the particle size from the amplitude of the scattered light, which is then converted to a mass concentration basis. The setup at the measurement site used a US EPA approved PM_{10} inlet. The sampling setup pulled in 16.67 L/min for the PM_{10} size cut, of which 5 L/min is the main flow to the instrument controlled by an internal pump, with the remaining 11.67 L/min being bypass flow maintained by an external pump. The T640x reported PM_{10} and $PM_{2.5}$ data at 1-min resolution.

3. Measurement Setup

95 and 89 PA-II sensor units were collocated with $PM_{2.5}$ FEM instruments for a 2 to 4-week period between December 2019 and January 2020 at Raleigh and Delhi. The FEM instruments stored the data on the instrument locally. The PA-II sensors transmitted data in real-time to the PA cloud server, from which the data were downloaded using an API. The PA data were available at 2-minute intervals. Figure 1a shows the schematic of the sample collection. The experimental setup at each location differed slightly and is described below.

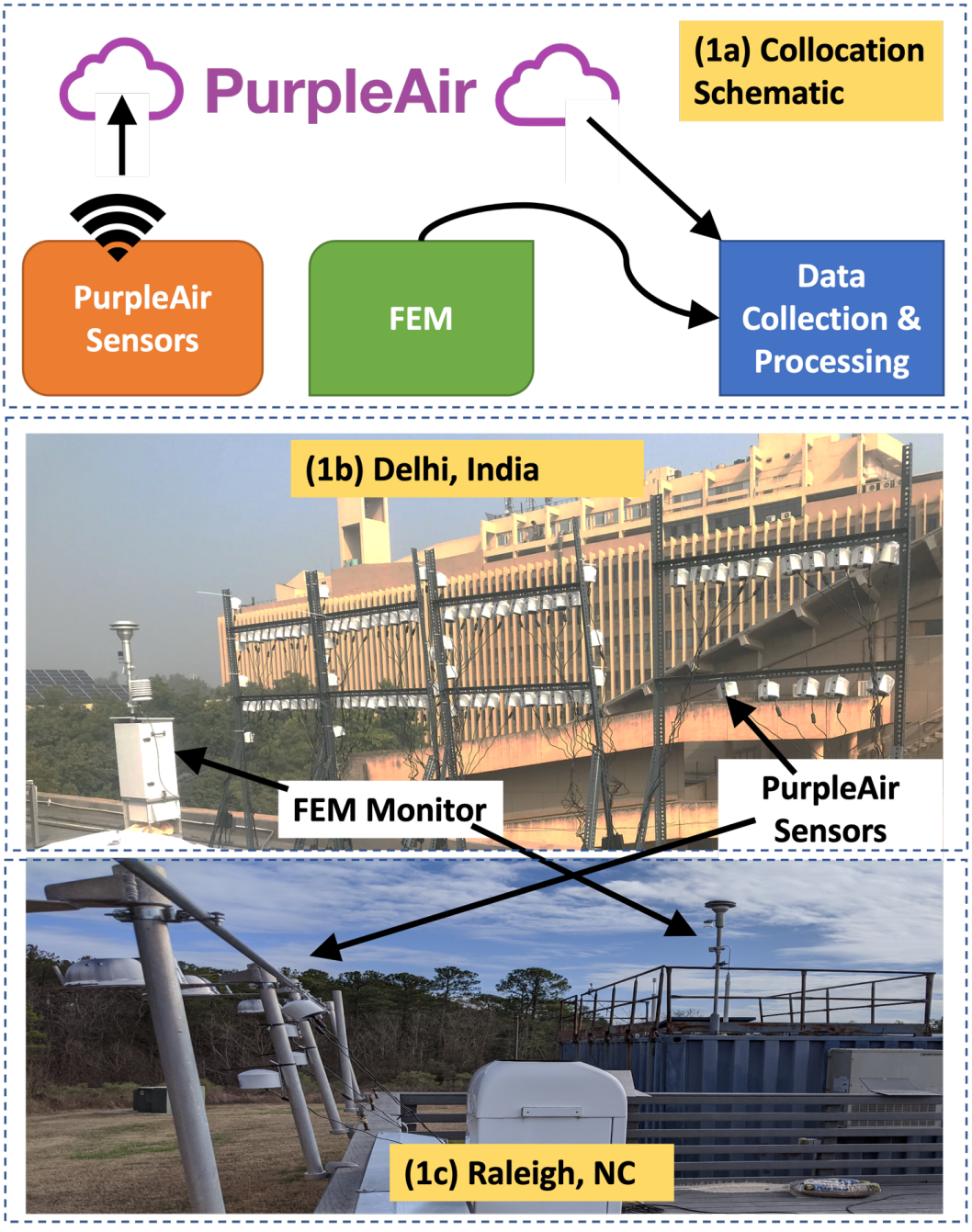
3.1. New Delhi, India

A total of 95 PA-II sensors were collocated next to the Met One BAM-1022 $PM_{2.5}$ FEM on a building rooftop at the Indian Institute of Technology, Delhi (IIT Delhi) (28.54464 °N, 77.19161 °E), New Delhi, India (Figure 1b) from December 19, 2019, to January 18, 2020. The PA-II sensors were mounted on a custom-built slotted metal stand about 2.74 m (~9 ft) tall. About 20 to 27 sensors were mounted using zip ties on each stand about 1.83 to 2.74 m (~6-9 ft) from the ground. The mounted sensors were roughly within 0.90 m (~3 ft) of the $PM_{2.5}$ inlet to the Met One BAM in the vertical direction and about 1 to 6 m (~3 to 20 ft) horizontally apart from the BAM. Ten battery-powered wireless hotspot routers were set up to generate wireless internet connections for real-time data transmission. Around 8 to 10 sensors were configured to use the Wi-Fi signal from any single hotspot as per specifications recommended by the manufacturer. The hotspot router battery was rated to have 7-8 hours of continuous operation time. All the PA sensors and the wireless routers were plugged into electrical outlets using extension cords with on/off control. The routers were connected to separate extension cords with dedicated on/off control. This allowed

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

the power to the sensors and the routers to be controlled separately. The routers were fully charged prior to use. To avoid potential battery damage and fire risk, the power to the routers was turned off once fully charged and manually turned back on after about 8 hours. Setting up a collocation of 95 sensors along with Wi-Fi hotspot units was a challenge. Due to issues with



the router and/or the extension cords, data were either unavailable or largely missing for 40 sensors. Thus, data from the remaining 55 sensors were used in this study.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022


Atmospheric Measurement Techniques Discussions

Figure 1: Generic data collection and calibration setup with actual pictures from the two locations. 1a) the schematic, 1b) the setup in Delhi, India, and 1c) the setup in Raleigh area, NC, USA.

3.2. Raleigh Area, NC

In the Raleigh area, the collocation was performed in two steps due to limitations in access to power at the collocation site. The first step involved the colocation of several PA-II sensors with each other in batches, sampling ambient air. A total of 89 sensors were collocated next to each other with some overlapping time periods at a residential location in Apex, NC, between December 16, 2019, and January 15, 2020. The sensors were inside a screened porch exposed to the ambient air. Five PA-II units were set as baseline units that were part of all the batches.

The second step involved collocating these five baseline units in the field next to the Teledyne T640x $PM_{2.5}$ FEM instrument at the ambient monitoring station at US E.PA (35.88952 °N, 78.874609 °W) in Research Triangle Park, NC, between January 16 to 27, 2020. At the field site, the FEM instrument was located inside a climate-controlled shelter with the sampling inlet extending through the roof (Figure 1c). The sensors were roughly 10-13 m away from the sensor inlet in the horizontal direction and within about 4.5 m of the FEM inlet in the vertical direction. All the sensors were connected to electrical outlets and configured to transmit data through Wi-Fi at the site. Data from the FEM were stored in an on-site data logger. Since the different steps were performed in different towns in the Raleigh region, we collectively refer to it as the Raleigh area or just Raleigh.

4. Data and Method

4.1. Data Integration and Evaluation Metrics

The PA-II $PM_{2.5}$ measurements from channels A and B were averaged at an hourly time interval to match the frequency of FEM measurements. Consistent data from the two channels were used as an indicator of sensor health, and data were preprocessed as discussed in Section 5.1. The preprocessed hourly average concentrations from the two channels were averaged together to obtain a single hourly average value for the PA-II sensor.

Hourly averaged $PM_{2.5}$ data from the FEMs, along with the average PA measurements of $PM_{2.5}$, T, and RH, were integrated into a single dataset separately for each location. This integrated and quality-controlled dataset, containing 18067 data records for Delhi and 1652 data records for Raleigh, was used to train and validate the calibration models.

To analyze the model calibration results and uncertainties, we considered the following statistical parameters:

$$Root\ Mean\ Square\ Error\ (RMSE), \mu g\ m^{-3} = \sqrt{\frac{1}{N}\sum(PM2.5_{FEM} - PM2.5_{PA})^2} \quad (1)$$

$$\text{Mean Bias (MB)}, \mu g\ m^{-3} = \frac{1}{N}\sum(\text{PM2.5}_{\text{PA}} - \text{PM2.5}_{\text{FEM}}) \quad (2)$$

The mean percentage bias (MB %) and the mean absolute percentage bias (|MB|%) of calibrated PA data are defined as:

$$MB(\%) = \frac{1}{N}\sum 100 * \frac{(\text{PM2.5}_{\text{PA}} - \text{PM2.5}_{\text{FEM}})}{\text{PM2.5}_{\text{FEM}}} \quad (3)$$

$$|MB|(\%) = \frac{1}{N}\sum 100 * \frac{(|\text{PM2.5}_{\text{PA}} - \text{PM2.5}_{\text{FEM}}|)}{\text{PM2.5}_{\text{FEM}}} \quad (4)$$

Where the $PM2.5_{FEM}$ is $PM_{2.5}$ from FEM monitors, $PM2.5_{PA}$ is the $PM_{2.5}$ measured by the PA-II sensor, and N is the number of paired data points. In addition, we have computed linear regression statistics, including Pearson correlation coefficient (R), slope (m), and intercept (I).

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022


Atmospheric Measurement Techniques Discussions

4.2. Machine Learning Algorithms for Calibration

Several models, including linear regression and selected MLAs, namely Support Vector Regression (SVR), XGboost, and Random Forest (RF), were tested. We looked at the RMSE, R, and MB metrics for these different models and chose the model with the lowest RMSE and a similar or better R than the other models. Based on the performance (Table 1), we selected RF as a candidate MLA for the detailed analysis in this study.

We used Scikit-learn (sklearn) machine learning library in Python (https://scikit-learn.org). The selected RF algorithm is a supervised MLA and one of the most used in modeling air quality using satellite remote sensing data sets (Masih, 2019) due to its simplicity and diversity. It randomly samples a small subset from the dataset and uses this to train multiple decision trees using the bagging method. The bagging method allows the combination of various learning methods, which improves overall accuracy. The ensemble of decision trees (i.e., forest) is then used to produce the final output. For the details on MLA parameter settings, we have provided a sample code for the training and testing of RF in the supplementary material.

In order to develop the final sensor calibration algorithm, we used the following steps: 1) quality control the PA data; 2) randomly divide the data into training (75%) and validation (25%) datasets; 3) train the algorithm using the training dataset and validate using the validation dataset; and 4) repeat steps 2 and 3 ten times (i.e., 10-fold cross-validation) using random data selection.

5. Results

The overarching goal of this paper is to report various aspects of low-cost sensor (i.e., PA-II) measurements and their calibration against FEM instruments. The results are presented side-by-side for the two regions and at different time averages to compare performance across regions. We present below an analysis of the difference in measurements between the two channels of a PA-II unit, differences among the multiple PA-II units, and a comparison of PA-II and FEM. Next, we present data preparation for calibration model development and testing of various MLA, including 10-fold validation. The last section presents the results of MLA and prognostic and diagnostic errors.

5.1. PA Sensor Data Quality Control

The quality of the data from each sensor must be checked prior to use in data analysis. In our earlier work (Gupta et al., 2018), we found differences in $PM_{2.5}$ values between these two channels for a few sensors. Figure 2 presents the scatter plots between PA-II channel A (PAA.) and channel B (PAB.) data for Delhi (top) and Raleigh (bottom) for hourly average concentrations. The different colors represent the different sensor units. Clearly, the two channels are highly correlated (R>0.95), and measurement values are close to each other, but there is also a spread in the scatter showing differences in values between the two channels. Overall, the mean (±one standard deviation) difference is 7±23% and 13±19% between the two channels in Delhi and Raleigh, respectively.

An initial examination of the sensor data indicated significantly different concentration regimes for the two regions of collocation (Figure 2). Delhi experienced hourly $PM_{2.5}$ ranging from about 50 to nearly 800 $\mu g\ m^{-3,}$ whereas Raleigh showed concentrations reaching a maximum of around 120 $\mu g\ m^{-3}$ but most often ranging between near zero and 25 $\mu g\ m^{-3}$ as measured by the PA-II. Due to the large difference in the range of $PM_{2.5}$ values, a single cleaning criterion will not work. As proposed in earlier studies (Barkjohn et al., 2021), criteria based on both percent difference as well as the absolute difference between

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

channels A and B are used based on the concentration regime to ensure high data quality for calibration model development. Consequently, we adopted separate data cleaning criteria for the two regions:

a) For Raleigh, where concentrations were typically less than 25 µg m^{-3}, data were excluded if the difference in hourly average $PM_{2.5}$ measurements between channel A (PAA.) and channel B (PAB.) was larger than 5 µg m^{-3}
b) For Delhi with higher concentration regimes, the data were excluded if the difference in hourly $PM_{2.5}$ between PAA. and PAB. was larger than 5%

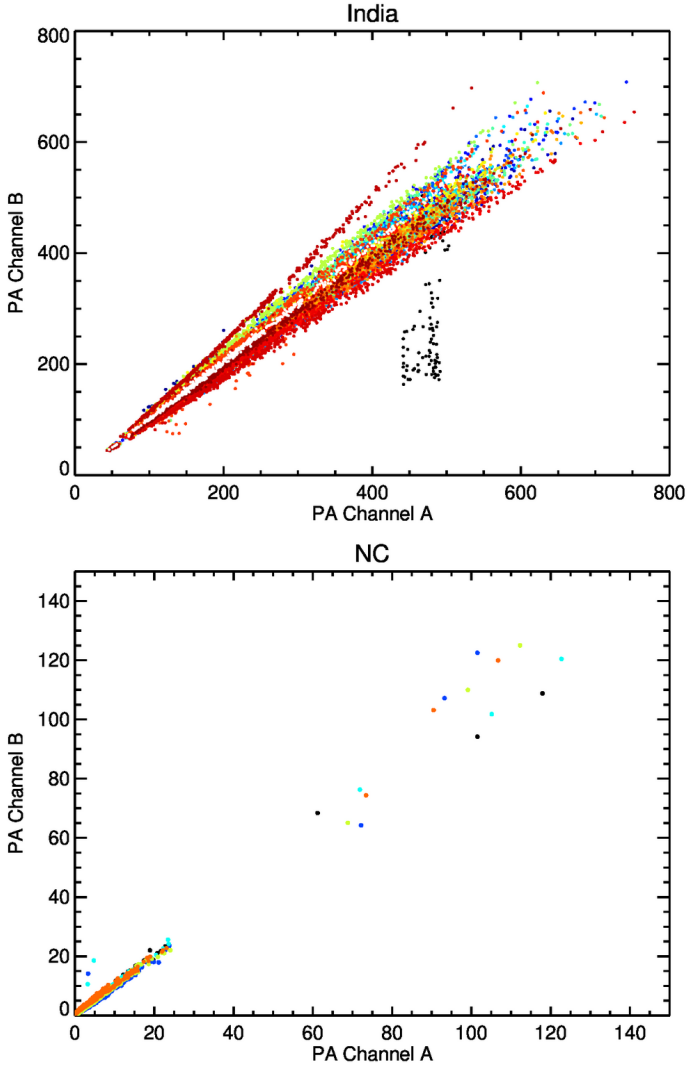


Figure 2: The scatter plots showing PA channel A & B raw measurements for the two regions (Top – Delhi, Bottom – Raleigh). The different color shows different sensor units.

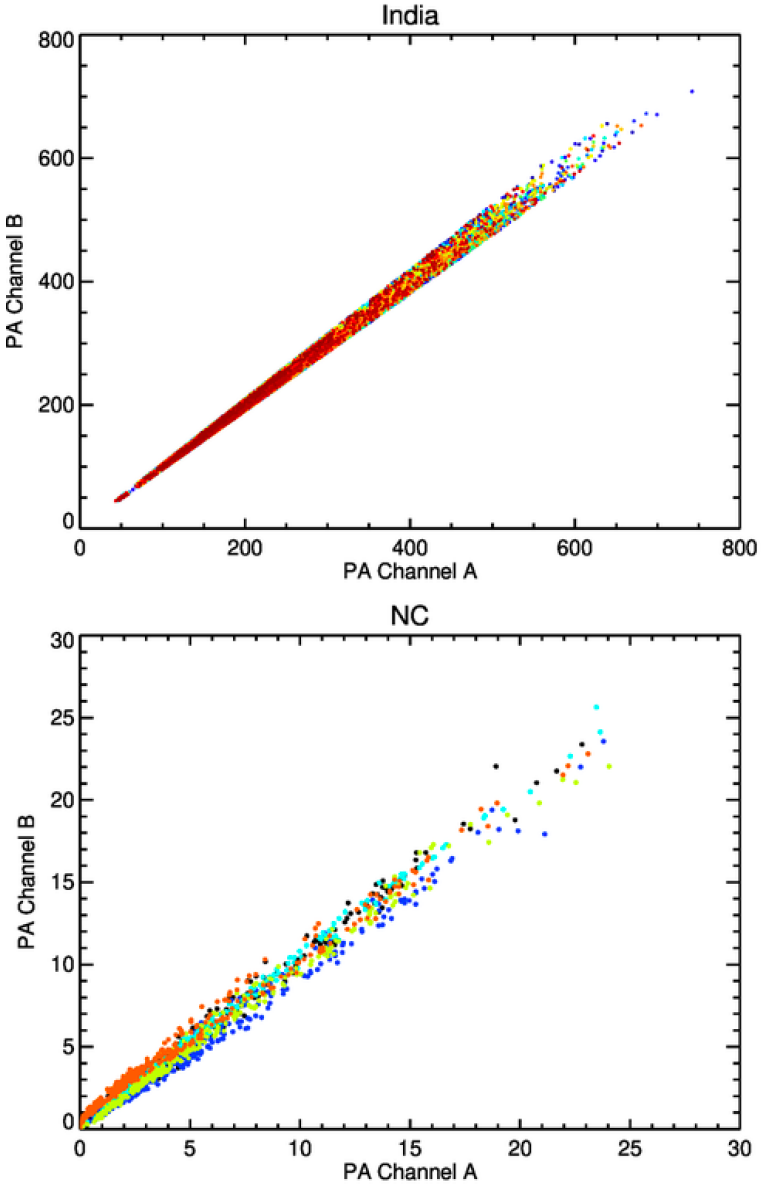
In addition, for Raleigh, we also removed a specific instance of high PA-II $PM_{2.5}$ values (> 50 µg m^{-3}) compared to FEM (6-8 µg m^{-3}) on January 12 and 13, 2019, between 11 pm and midnight (total of 18 data points). Although the discrepancy with FEM itself should not be a reason to discard the data, the fact that there were no known explainable reasons and that this was

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

an isolated instance during the nearly 2-week collocation period, we decided to remove these data as well. The main goal is to retain only the highest quality data that would help identify and capture the underlying mechanism influencing LCS response compared to a FEM. We, therefore, deliberately chose stringent criteria to remove any potentially erroneous data or discrepancies that might mask the sensor-FEM relationship. It is important to note that the data cleaning criteria can vary depending on the purpose of data usage and the level of tolerance for the errors.

PA-II data were cleaned and filtered out as per the criteria discussed above. Figure 3 shows the PA-II data after the application of quality checks, showing an excellent correlation (>0.95) between PAA. and PAB. with minimal spread around the 1:1 line and no outliers. We use the cleaned PA-II data for further analyses and model development.



https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022


Atmospheric Measurement Techniques Discussions

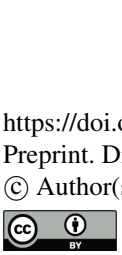
Figure 3: Same as Figure 2, except data quality is controlled and cleaned up using criteria discussed in section 5.1

Figure S1 (supplemental material) shows the mean (and standard deviation) of the difference between PAA. and PAB. for each sensor after data cleaning. The mean difference varied between -8.80 ± 4.55 μg m^{-3} and 9.97 ± 5.90 μg m^{-3} for Delhi and between -0.64 ± 0.42 μg m^{-3} and 0.31 ± 0.50 μg m^{-3} for Raleigh. When examining data between any two sensors, the mean difference among the PA-II sensors ranged from -33.3 ± 18.7 μg m^{-3} to 70.3 ± 24.5 μg m^{-3} for Delhi (Figure S2.1) and from -0.29 ± 0.35 μg m^{-3} to 0.59 ± 0.23 μg m^{-3} for Raleigh (Figure S2.3). When expressed as a percent of the mean of the two sensors, the sensors were within a maximum difference of ± 20% of each other in Delhi (Figure S2.2). Out of a total of 1262 valid sensor pair combinations, about 77% of the pairs showed differences within 5% of each other and 96% of the pairs within 10% of each other. For Raleigh, on a percent basis, the sensors differed between -9.2% and 23% (Figure S2.4). Out of a total of 10 valid sensor pair combinations, about 20% of the pairs were within 5% of each other and 50% of the pairs within 10% of each other. The higher percent difference between sensors in Raleigh is due to division by a low concentration range because the absolute difference was within 0.54 μg m^{-3} for 90% of the sensor pairs.

5.2. PA and FEM Intercomparison

After we cleaned and quality-controlled the PA data, we compared the average of channel A (PAA.) and channel B (PAB.) with coincident hourly FEM $PM_{2.5}$ values. Figure 4 shows these comparisons at hourly (left) and daily (right) averages for Delhi (top) and Raleigh (bottom). From the figure, it is clear that the PA-II sensors have very different performances against FEM in the two regions, mainly due to differences in $PM_{2.5}$ loading, particle type, and operating conditions. The typical $PM_{2.5}$ values in Raleigh were less than 20 μg m^{-3} which were rarely observed in Delhi. The PA vs. FEM comparison in Raleigh showed a big scatter in hourly averages with a R value of 0.34 and RMSE of 4.4 μg m^{-3}. After averaging data over a 24-hour period (i.e., daily average), the correlation almost doubled (R = 0.66), and RMSE was reduced by half (2.2 μg m^{-3}). The mean bias for Raleigh remained negative and about the same (~ -0.8 μg m^{-3} or about 5% to 11%) on both hourly and daily average basis, suggesting an overall underestimation by PA in clean conditions ($PM_{2.5}$ < 10 μg m^{-3}). In contrast, the PA sensors in Delhi often overestimated $PM_{2.5}$ concentrations at both hourly and daily averages with a positive mean bias (~35 to 37 μg m^{-3}, or 23.8% to 23.7%). PA measurements in Delhi showed a very high degree of correlation (R ≥0.88) with FEM but with a high RMSE of 60.75 μg m^{-3} and 48.13 μg m^{-3} on an hourly and daily basis, respectively. Thus, the comparison of PA with FEM demonstrated completely different sensor behavior for Delhi (overestimation but highly correlated) and Raleigh (underestimation and low correlation). This suggests the need for different calibration coefficients or models for correcting PA data under different $PM_{2.5}$ loadings. It is also important to note that the chemical composition of particles in Delhi and Raleigh is expected to be different. The Delhi particles are dominated by carbonaceous aerosols with a mixture of dust during the winter period (Shiva Nagendra and Khare, 2019), whereas PM in Raleigh is dominated by typical urban sulfate and nitrate aerosols (Cheng and Wang-Li, 2019).

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022



Atmospheric Measurement Techniques Discussions

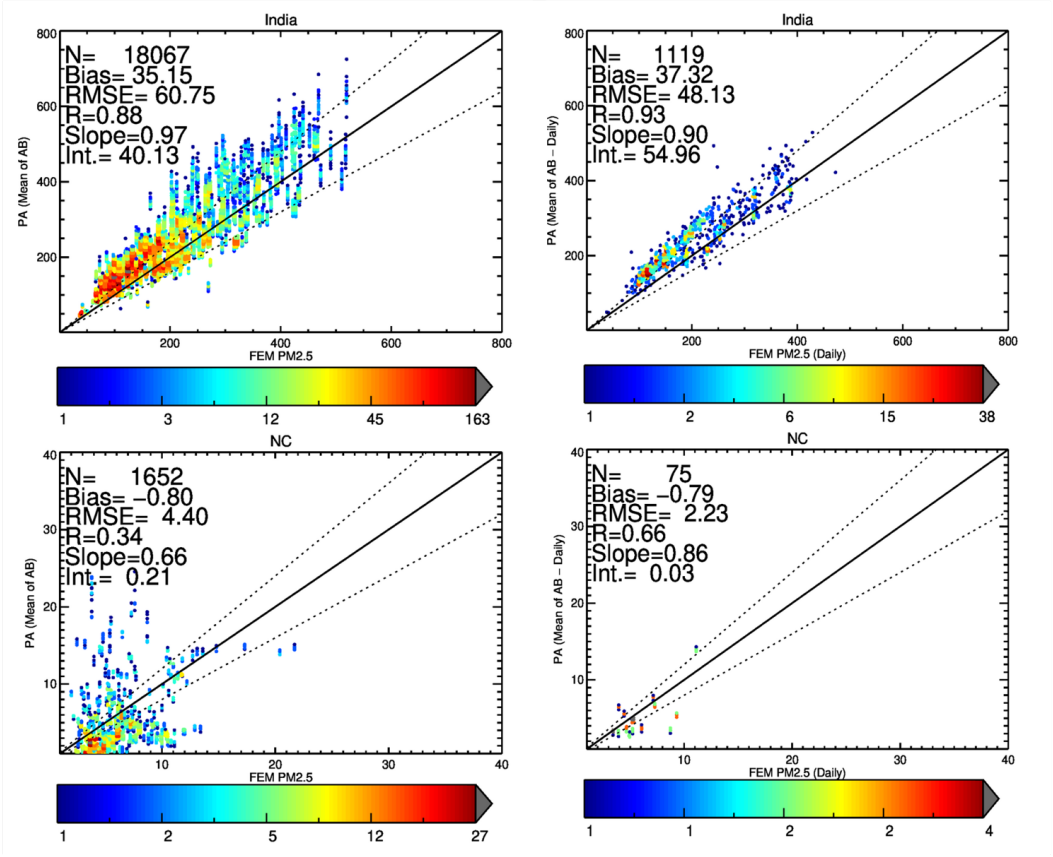


Figure 4: The density scatters plot of FEM and PA measurements of PM2.5 for the two locations after data cleaning for hourly (left) and daily (right) mean values. The color scale represents the density of data points.

5.3 PA Calibration Algorithm Performance

Our goal is to develop a robust calibration model which uses PA measured parameters (i.e., $PM_{2.5}$, T, RH) as inputs and generates bias-corrected $PM_{2.5}$ equivalent to that measured by a FEM as an output. As discussed in section 4.2, several MLAs have been trained and tested using integrated PA-II and FEM datasets (section 4.1). Table 1 presents the results of model performance (RMSE, R, and MB) during training and validation (testing) using linear regression, S.V.R., XGBoost, and RF algorithms for both Delhi and Raleigh regions. Almost all the performance metrics indicated superior performance by the RF algorithm compared to other methods. The RF model showed the lowest RMSE, highest R, and lowest bias for both regions during the training phase. Similar performance was observed for the validation phase for Delhi. For Raleigh, even though mean bias was not the lowest during the training phase, the RF approach still yielded the lowest RMSE and the highest R with an overall best performance when considering all metrics in tandem. The performance of the RF model was slightly degraded during the validation phase compared to the training phase for the Raleigh region, likely indicative of the higher uncertainty associated with model calibration at very low concentration regimes. This could be due to low variability in input and output parameters across the data distribution in Raleigh. Thus, based on the initial testing and performance of multiple MLAs, as presented in table 1, we selected the RF method for further analysis and calibration algorithm (or model) development.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

Table 1: The summary statistics of ML algorithms performance for training and testing datasets.

Raleigh, NC	Training N = 1239			Testing N = 413		
ML Algorithm	RMSE	R	Mean Bias %	RMSE	R	Mean Bias %
Linear Regression	2.7	0.12	17.8	2.6	0.12	20.1
SVR	2.6	0.13	7.1	3.0	0.09	4.0
XGBoost	1.7	0.62	5.5	2.0	0.57	4.6
Random Forest	**0.7**	**0.97**	**3.1**	**1.7**	**0.85**	**7.6**

Delhi, India	Training N=13,550			Testing N=4517		
ML Algorithm	RMSE	R	Mean Bias %	RMSE	R	Mean Bias %
Liner Regression	41.3	0.82	17.5	40.7	0.81	17.9
SVR	49.9	0.75	19.8	49.4	0.74	20.2
XGBoost	33.2	0.88	12.8	33.5	0.88	13.6
Random Forest	**20.4**	**0.98**	**1.7**	**30.3**	**0.85**	**2.8**

Figure 5 presents the performance of the final RF model (average (i.e., Ensembled) of the 10-fold cross-validation) for both Delhi and Raleigh. The top panels show performance during training and validation for Delhi, while the bottom panels show the performance for Raleigh. The comparisons for both regions clearly show that the training data performed slightly better than validation data across the different statistical parameters. The Delhi model demonstrated high correlation ($R \geq 0.95$), low mean bias (< 1 µg m^{-3}), slope close to 1 (> 0.95), and RMSE of about 20 to 30 µg $m^{-3,}$ respectively, for training and validation datasets. The Raleigh model performance showed low bias (< 0.1 µg m^{-3}) and RMSE of 0.69 to 1.7 µg m^{-3} and a high correlation of 0.97 to 0.85 but slopes less than unity (0.89 to 0.69) for training and validation datasets respectively. The large difference in $PM_{2.5}$ loadings over the two regions makes it a complex problem to model together, requiring two different calibration models and likely the reason for the slightly lower performance for Raleigh. Figures S3.1 to S3.4 in the supplemental material show the 10-fold training and validation performance metrics during the ML model development process for the two regions. The density scatter plots in Figure S3 show consistent results across the 10-fold simulations for training and validation steps confirming the optimized nature of the selected calibration model.

As noted in section 3.2, for the Raleigh area, only 5 PA sensors were directly collocated with FEM, and the remaining 84 sensors were tested against those 5 sensors. In figure S4, we present the comparison of ML corrected 5 sensors mean with the

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

remaining 84 sensors after ML corrections. The high correlation (R=0.96) and slope value close to one (0.91) between sensors after ML corrections demonstrate consistent model behavior.

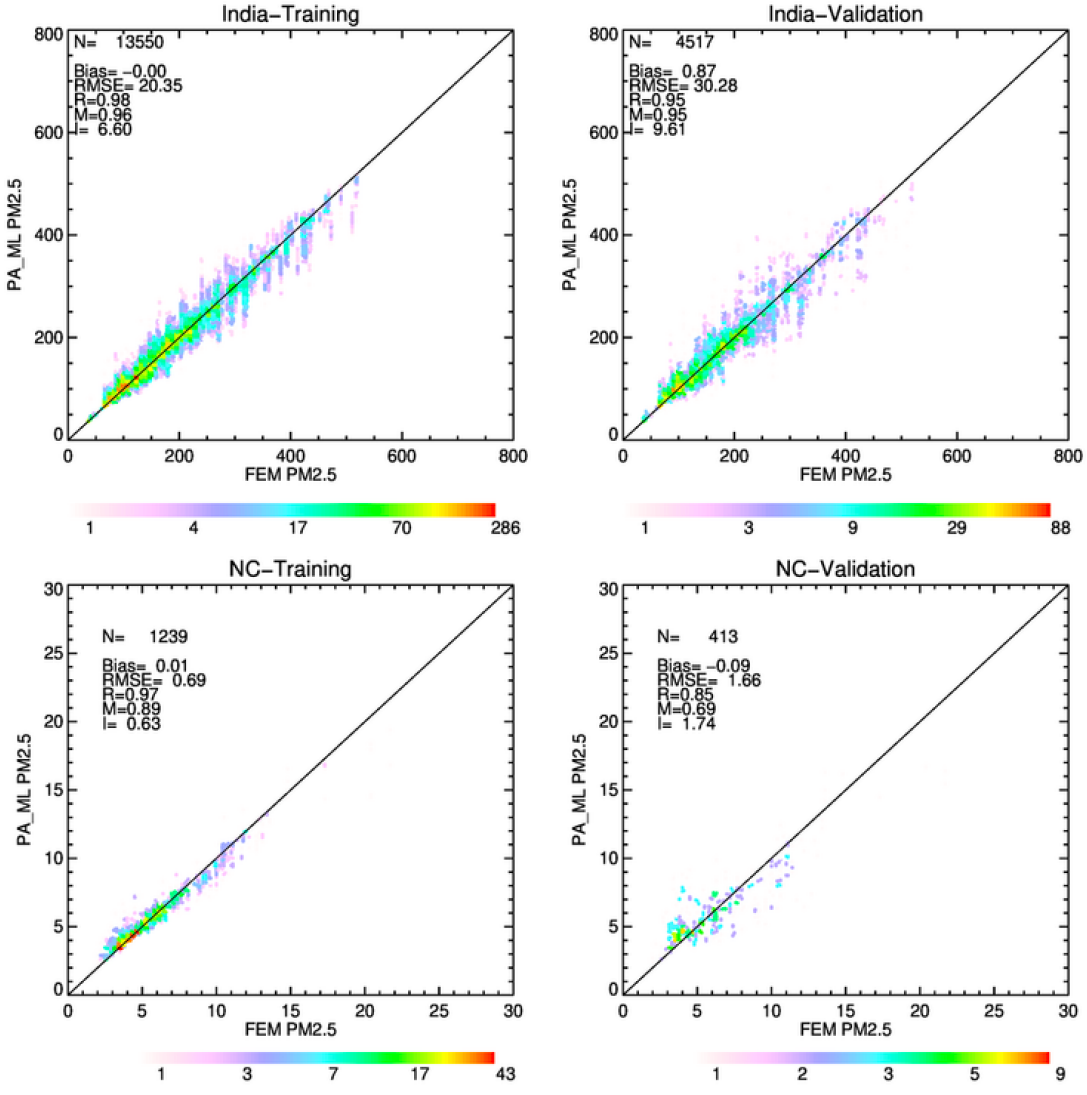


Figure 5: The density scatter plots showing a comparison between FEM and output from MLA. The top panels are for Delhi training and testing of MLA, respectively, while the bottom panels are for Raleigh. The colors represent the density of data points.

We also evaluated the importance of each input parameter and presented the results in Figure S5. Here, the importance of each input parameter in estimating FEM equivalent $PM_{2.5}$ is reported as a percentage. It is interesting to note that the relative importance of the parameter differed by region, likely due to the different concentration regimes and particle composition. For Delhi, with high concentration ranges, the PA $PM_{2.5}$ data was the major input parameter (~85% of the total score), with T and RH each contributing roughly about 7.5%. On the other hand, for Raleigh, PA, $PM_{2.5}$ ranked at ~48% score, with T contributing about ~29% to the importance score, followed by RH at 23%. Therefore, under low concentration settings such as those observed in Raleigh, the importance of meteorological variables combined was similar to or greater than that of $PM_{2.5}$ data.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

Although PA sensors provide 2-minute resolution data, our analysis is focused on hourly and daily averages, which are typically reported by the FEM instruments. The hourly data are often used to estimate current or real-time air quality conditions (i.e., "NowCast" of the air quality index (AQI)), whereas US national ambient air quality standards for $PM_{2.5}$ are based on a 24-hour average. Thus, we evaluated the performance of MLA for both hourly and daily averages. Figure 6 shows the performance of calibrated PA measurements against the FEM on hourly (left plot) and daily (right plot) averages for the two regions. It is important to note that this figure contains all the points in the integrated datasets, unlike Figure 5, where training and validation data were presented separately. On a 24-hour average basis, the calibrated PA values showed excellent correlation ($R > 0.98$), low mean bias (1.3 and 0.03 $\mu g\ m^{-3}$ for Delhi and Raleigh, respectively), and slopes within 5 (Delhi) to 11% (Raleigh) of unity, demonstrating excellent performance for 24-hour average values. The mean absolute percentage bias (Eq. 4) for hourly and daily mean values are respectively, 9.1±9.8 and 5.4±6.3 for Delhi and 10.9±16.0 and 5.0±5.3 for Raleigh. The mean $PM_{2.5}$ concentration of Delhi and Raleigh was 193 and 6 $\mu g\ m^{-3}$.

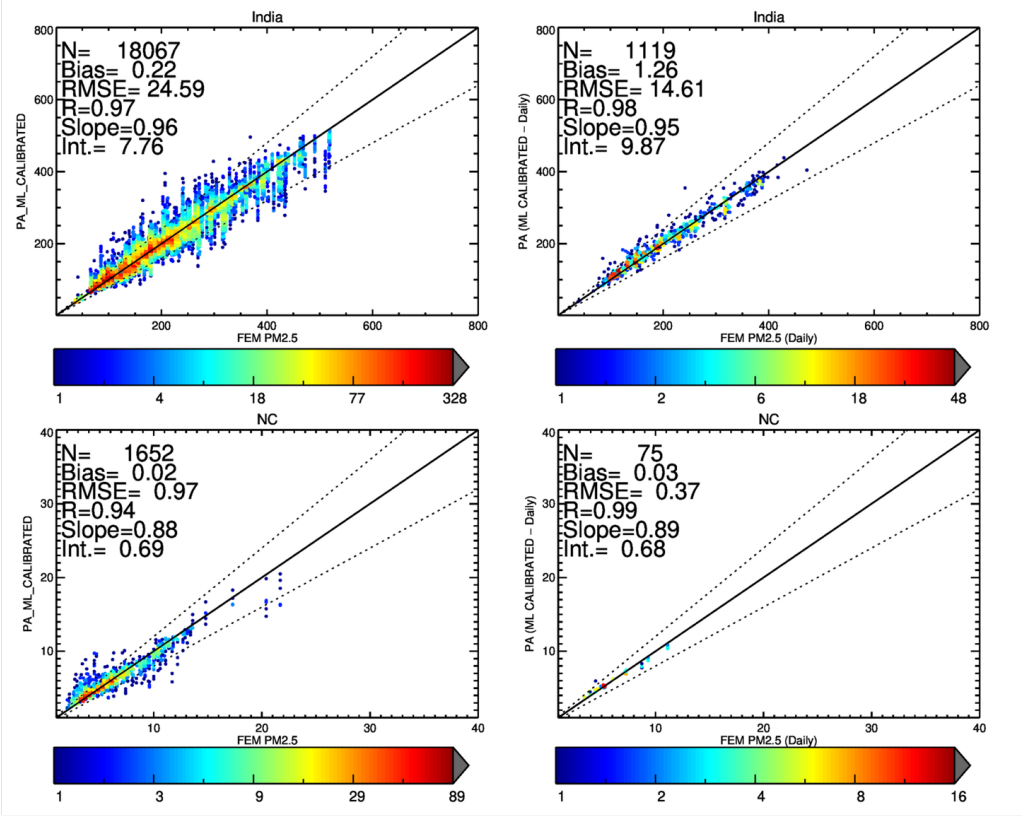


Figure 6: The density scatter plots showing a comparison between FEM and output from MLA for hourly (left panels) and daily mean (right panels). Here data for both training and testing are combined. The top panels are for Delhi, and the bottom panels are for Raleigh.

To further understand the biases, figure 7 shows the frequency distribution of hourly and daily average biases for all the data for the two regions. An analysis of the bias distribution shows that about 67% and 94% of hourly data fall within ±10% and ±25% biases for Delhi, while for Raleigh, about 68% and 90% of the data were within ±10% and ±25% biases. The bias distribution for 24-hour average values shows significant improvement with a similar percentage of data (62% and 94% for

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

Delhi and 64% and 96% for Raleigh) falling within smaller bounds of ±5% and ±15% bias. In other words, on an hourly basis, about 90 to 94% of the data fell within ±25% bias, while on a daily average basis, a similar proportion of the data (94 to 96%) fell within a lower bias of ±15% demonstrating improved model performance for daily averages.

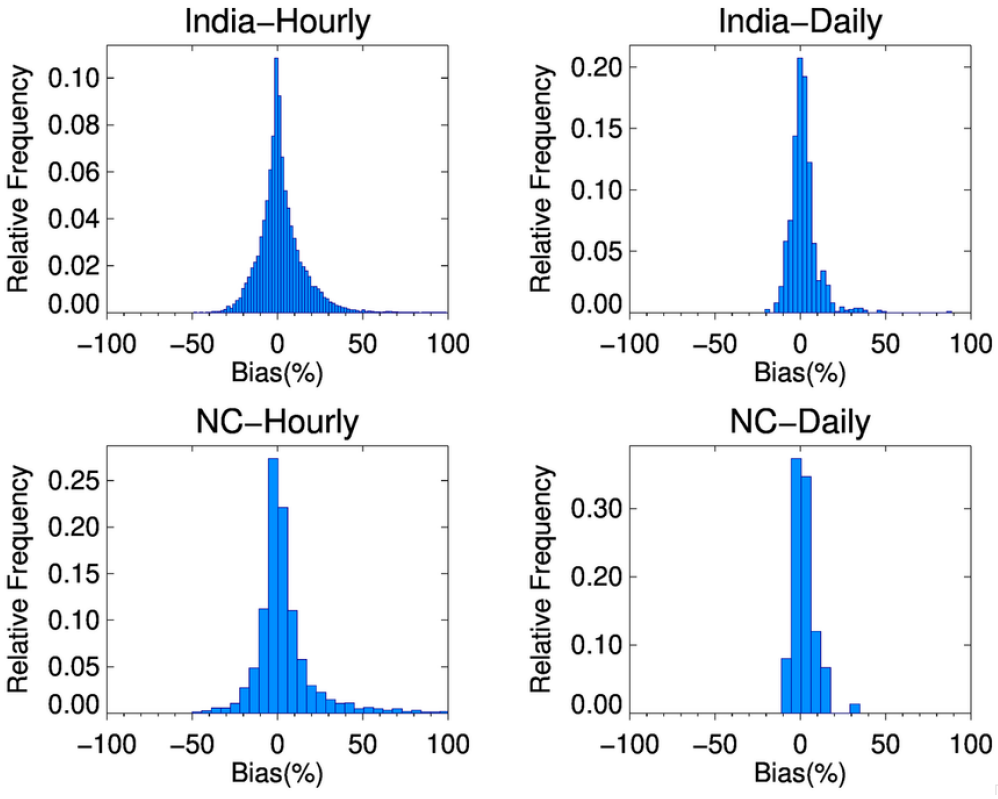


Figure 7: Frequency distribution of MB (%) for each region (hourly and daily data).

Figures 8 (Delhi) and 9 (Raleigh) demonstrate the capability of PA to track diurnal and day-to-day changes in $PM_{2.5}$ concentration as compared to FEM measurements. The top panels show the raw PA data for hourly (left) and daily (right) averages, whereas the bottom panels display the MLA calibrated data. The red line represents the mean PA data, with the shaded area representing one standard deviation among multiple PA sensors tested at each location. In Delhi, the raw PA data followed the FEM in both hourly and daily trends with an overall positive bias, whereas raw PA data in Raleigh demonstrated more random variability for hourly data compared to FEM. The daily PA data followed FEM but with an overall random negative bias for Raleigh. After MLA calibration, PA data in both regions followed FEM very nicely with minimal deviation. This analysis demonstrates the temporal consistency of calibrated PA data and its application for monitoring both diurnal and day-to-day variability.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

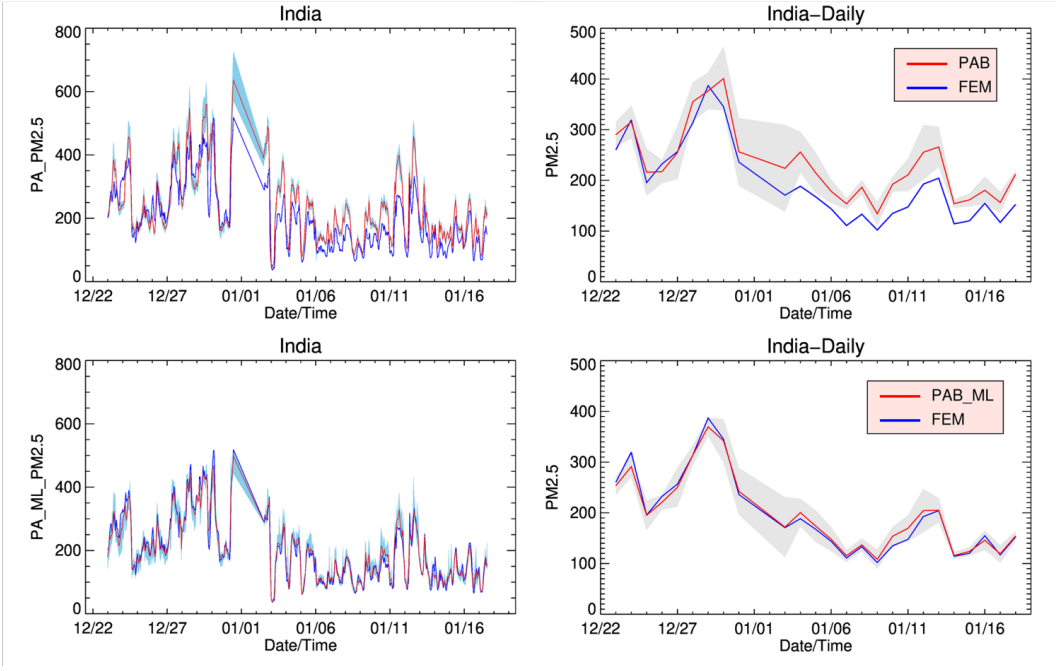


Figure 8. The hourly (left) and daily (right) variation in $PM_{2.5}$ using PA (raw, top panels) and ML-corrected PA (bottom panel) compared to FEM for Delhi.

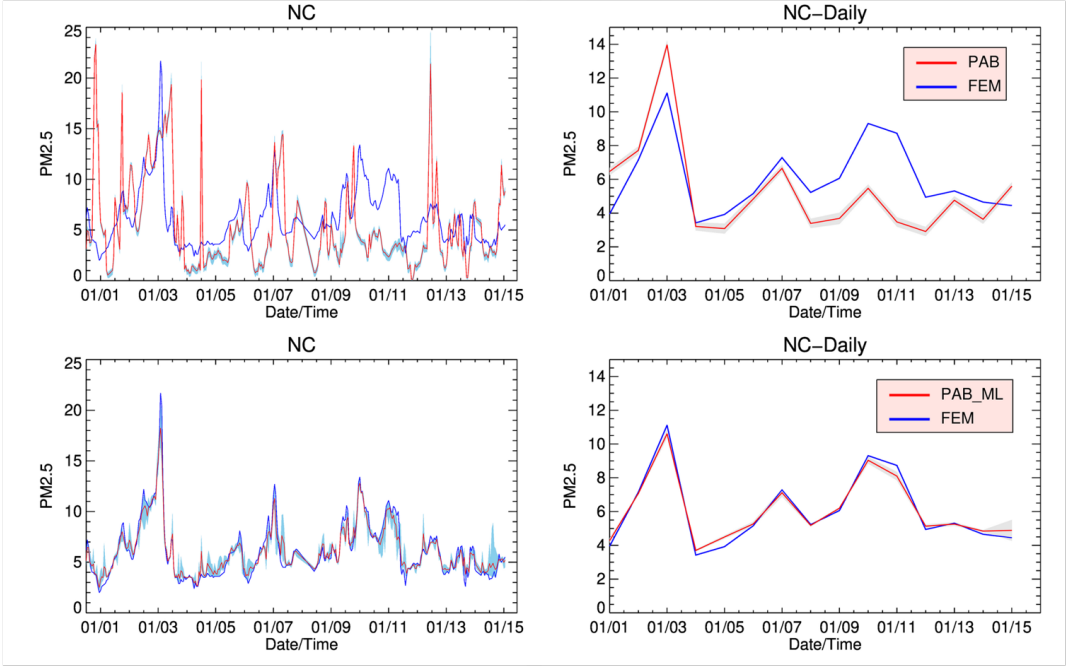


Figure 9. Same as Figure 8 except for Raleigh.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

5.4 Error Characterizations

We next explore the relationship between the PA vs. FEM $PM_{2.5}$ bias and the various input parameters to MLA. At each coincident hourly pair of PA and FEM, the FEM $PM_{2.5}$ was subtracted from the PA $PM_{2.5}$ so that a positive difference indicates a positive PA bias (i.e., overestimation) and a negative bias represents underestimation by PA. The data was then sorted according to a parameter of interest in the database and repeated for both raw and calibrated PA datasets. Collocations were grouped into 17 bins for Raleigh and 37 bins for Delhi, each bin containing 100 and 500 pairs, respectively. Thus, there were equal numbers of PA-FEM pairs in each bin, but the bins were not equally spaced along the x-axis. The mean, median, and standard deviations of the PA-FEM differences were calculated for each bin.

Figure 10 (a-d) shows the results of this analysis for Raleigh for bias in raw PA data as a function of FEM $PM_{2.5}$, raw PA $PM_{2.5}$, T (TEMP, °F), and RH (%), while figure 10(e-h) shows similar comparisons but for MLA-calibrated PA data. The red and green colors show bin mean and median values, while the shaded color represents one standard deviation. PA raw data showed (Fig. 10a) positive biases for very low FEM $PM_{2.5}$ (< 5 µg m^{-3}), while biases were negative for the rest of the $PM_{2.5}$ concentration ranges. As expected, the biases as a function of raw PA $PM_{2.5}$ (Fig 10b) were reversed (i.e., negative for low $PM_{2.5}$ values and positive for high $PM_{2.5}$ values). We plot these differences against the PA-measured $PM_{2.5}$ to create a metric of accuracy that can be used to evaluate individual PA measurements. The biases remained negative as a function of temperature (Fig 10c) and did not demonstrate any specific pattern but appeared to be more random. Similarly, biases with respect to change in RH (Fig 10d) also oscillated between negative and positive values. After MLA calibration was applied to the PA datasets, biases remained flat near zero with a narrower standard deviation for all the dependent parameters (Fig 10e-h).

Figure 11 shows the same analysis for Delhi and demonstrates similar results with some differences, as noted here. The biases in raw PA $PM_{2.5}$ as a function of FEM $PM_{2.5}$ (Fig 11a) remained consistently positive with some random variability. The biases also remained positive with raw PA $PM_{2.5}$ and increased with an increase in $PM_{2.5}$ values (Fig 11b). It is important to note that the range of temperature and RH over Delhi was larger compared to that in Raleigh during the measurement period. The biases appeared to be decreasing as a function of temperature (Fig 11c) and increasing as a function of RH (Fig 11d). The lower temperature was typically also associated with higher $PM_{2.5}$ values in Delhi, which can explain the dependence of biases on temperature. The FEM instrument measured $PM_{2.5}$ at controlled RH (< 50%), whereas PA sensors measured at ambient conditions (30 to > 85%). The impact of RH on particle scattering property is well known (Adam et al., 2012) and can create biases in PA at high RH values. The bias in MLA calibrated PA $PM_{2.5}$ values (Fig 11e-h) as a function of the different parameters was flat near zero except for FEM $PM_{2.5,}$ where it demonstrated some negative bias when FEM $PM_{2.5}$ exceeded 300 µg m^{-3}. This analysis is important to understand the PA performance under different pollution levels and environmental conditions.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

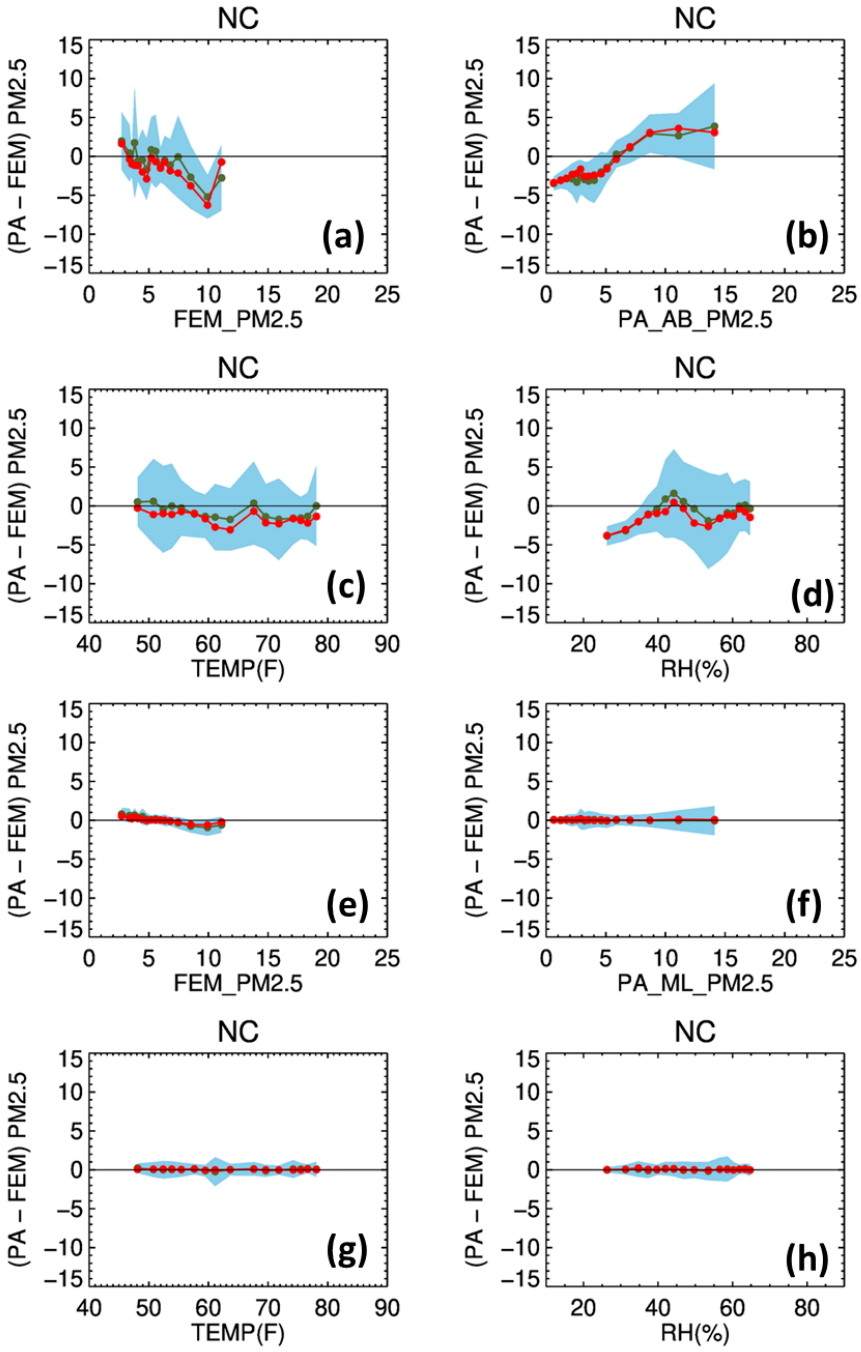


Figure 10: Diagnosis and prognosis errors in hourly data for Raleigh before and after ML calibrations.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

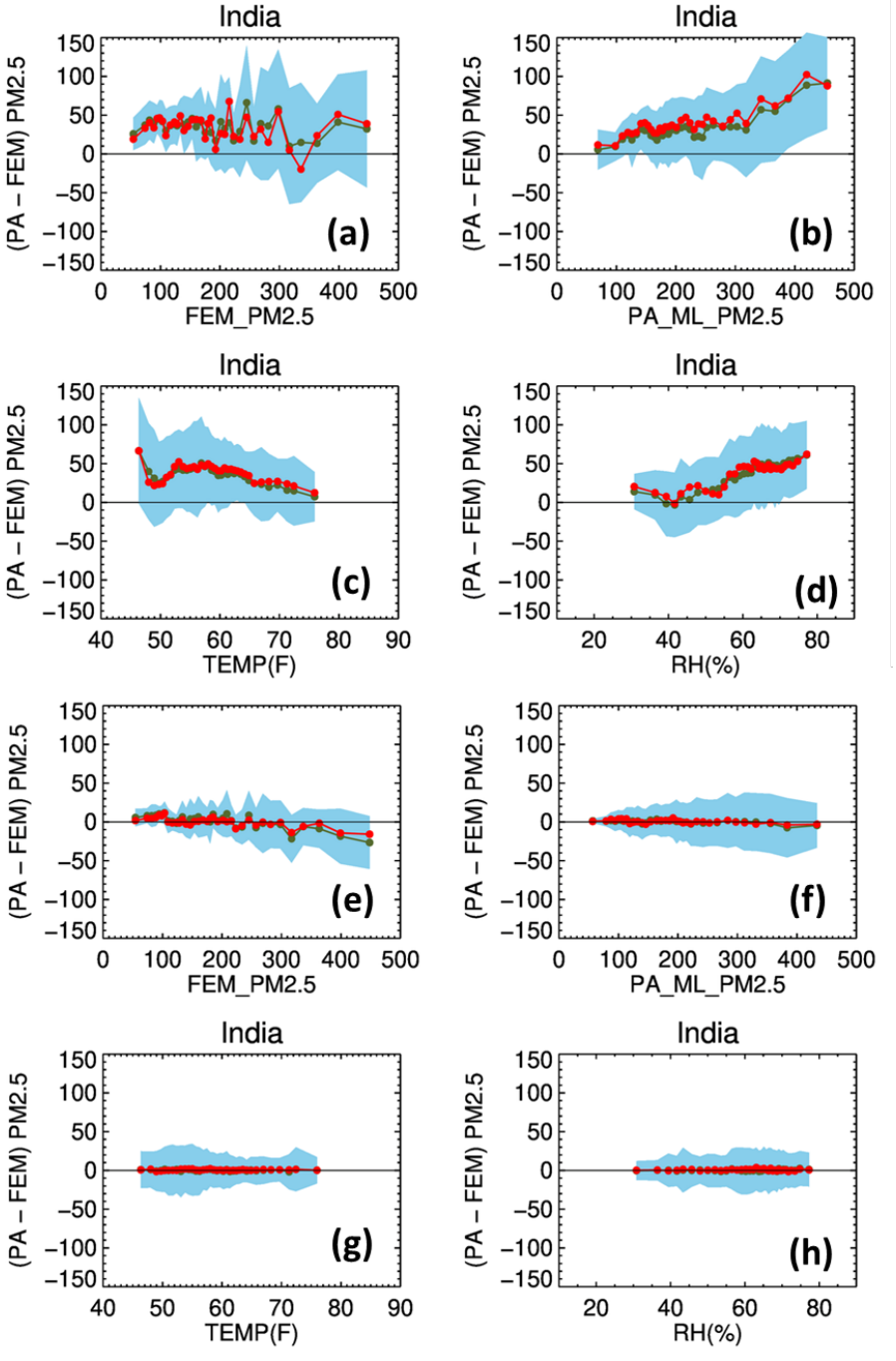


Figure 11: Diagnosis and prognosis errors in hourly data for India before and after ML calibrations.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Atmospheric Measurement Techniques Discussions

6 Discussion

Our study presents an evaluation of low-cost sensors in two different regions with contrasting aerosol characteristics, loadings, and environmental conditions. To our knowledge, this is the first study that performed a simultaneous evaluation of multiple sensor units on a large scale (>50 sensors) by collocating it next to a FEM over two different continents. The large database of sensor-FEM measurements allowed us to develop a robust correction model capturing the inter-sensor variability. The simultaneous collocation of multiple sensors also allowed us to estimate the uncertainty in the bias-corrected estimations by the model. Our results show that the final model improved the correlation, reduced the RMSE by more than 50%, and significantly reduced the error (Eq. 4) (from 29% to 9% in India and from 60% to 11% in Raleigh, NC) in MLA estimated hourly $PM_{2.5}$. Table S1 provides various bias estimations for raw and calibrated data for both hourly and daily averages for the two locations.

In an earlier study, Zheng et al. (2018) reported calibration of custom-made LCS unit using the Plantower PMS3003 sensors at the same location in Raleigh, NC, during summer (July 2017) and at another city in North India (Kanpur), about 400 km southeast of Delhi, India, along the Indo-Gangetic Plain during Oct-Nov 2017. The study used a linear or quadratic model to correct sensor data (RH-adjusted for periods with high RH influence) using the entire dataset (i.e., no holdout data for evaluation). They presented models with no RH adjustment for both hourly and daily averages and with RH adjustment and RH+T adjustment for the hourly data. For the Raleigh region, their model resulted in an R^2 of 0.66 (no RH adjustment) to 0.95 (RH adjustment and T correction) for hourly average values and 0.94 (no RH or T correction) for daily average concentrations. The ratio of calibrated sensor value to reference data ranged from 0.99±0.27 to 1±0.08 for 1-hour average to 1±0.09 for daily average concentration. For India, mean R^2 ranged from 0.61 to 0.78 with calibrated sensor to reference ratios of 0.96 to1.01 for hourly average and R^2 of 0.78 to 0.93 with ratios of 0.99 to 1 for daily average concentrations. Using metrics for the training period as an equivalent point of comparison, our study utilizing an MLA that incorporates RH and T effects yielded R^2 of 0.94 (Raleigh) to 0.96 (India) with a mean percentage bias of 4.2±18.9% (Raleigh, or a ratio (corrected/F.E.M.) 1.04±0.19) and 2.0±13.2% (India, or a ratio of 1.02±0.13) for the training portion of hourly data, demonstrating similar or better performance than the Zheng et al., (2018) findings.

In another study, Barkjohn et al., (2021) developed a US wide correction approach for PA sensors using a multiple linear regression model that included RH correction. They used PA sensors deployed in the field that were within 50 m of a FEM site. In that study, for the state of North Carolina, the model predictions resulted in a RMSE of 2.1 µg m^{-3} with a mean bias of 1 µg m^{-3} for a 24-hr average $PM_{2.5}$ at a site about 120 km away from the site used in our study. Magi et al. (2019) performed a similar evaluation of PA sensors against the FEM (BAM) at Charlotte, NC, about 200 km southwest of Raleigh, NC. They developed a multiple linear regression model that resulted in a RMSE of 4.1 µg m^{-3} with a correlation (R^2=0.6) for hourly $PM_{2.5}$. Our MLA for Raleigh (NC) yielded an improved RMSE of 1.7 µg m^{-3} and a mean bias of nearly zero (R^2=0.72) on the holdout sample for hourly $PM_{2.5}$. For the daily averages, the RMSE dropped to 0.37 µg m^{-3} with mean bias remaining near zero. The improved performance seen in our study might be due to the type of model used, coincident spatial collocation of PA units, and difference in observed $PM_{2.5}$ ranges. Our model is specific to this site, while the Barkjohn et al. (2021) model was optimized for U.S.-wide correction and applied to NC data. Further, our data was limited to winter, while the Barkjohn (2021) and Magi et al. (2019) studies covered nearly 16 to 18 months of data, which may be the reason for the larger variability seen in those studies.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022


Atmospheric Measurement Techniques Discussions

Our comprehensive collocation dataset allowed us to develop a robust model based on machine learning techniques for both regions that demonstrated similar or better performance than prior studies. However, our collocation was limited to one month at both locations due to the limited availability of resources involved in the collocation of multiple sensors at one location. Additional logistical challenges arose due to restrictions in place as a result of pandemic-related lockdowns ongoing at that time. The model is therefore likely optimized for the winter season (and the associated weather conditions) when the collocation was performed. The performance of the model for other seasons (and thus for other weather conditions) will need to be evaluated as part of future work. We have a couple of sensors deployed next to a FEM as part of a long-term collocation effort to study the seasonal differences in model parameters and other sensor performance metrics such as sensor drift. From our assessments and that published in the literature, it is known that the sensor performance will vary by aerosol composition and loading. Therefore, the applicability of the model is probably limited to regions with similar aerosol composition and source influence. The applicability of the model to other geographic regions has not been tested. Other minor shortcomings of our study include the lack of ideal sensor mounting, potential differences in the direction of airflow among sensors, and the distance and height of sensor mounting with respect to FEM instruments. Nevertheless, the ability of the models to generate bias-corrected data that are within about 5 to 10% of reference data on hourly and daily timescales, respectively. For context, the acceptable measurement uncertainty in FEMs is typically about 10% (US E.PA, 2016). Thus, the LCS offers great potential in generating high-quality data when corrected appropriately for biases. Such data offer promise for applications in air quality management, including understanding air pollution burden, the temporal and spatial characteristics, air quality forecasts, exposure assessments, and filling in gaps in regulatory monitoring data to support integrated datasets that may ultimately support policy. This is partly dependent on the validity of the calibration in a real-world deployment. Our future work will assess how well these models hold and perform for sensors deployed in the field away from the collocation site and for how long and examine the utility of sensor data for such applications that will provide insights into their practical utility and limitations.

7 Summary and Conclusion

In this study, we collected simultaneous measurements of $PM_{2.5}$ mass concentration using the PA-II LCS and FEM monitors in Raleigh, NC, and Delhi, India. In Delhi, we had data from 51 PA sensors, whereas, in Raleigh, 5 PA sensors were deployed next to a FEM, and additional 85 sensors were collocated with these five sensors in batches. The coincident measurements from PA sensors and FEM monitors from both locations were quality controlled and used to assess the performance of PA sensors against those of FEMs. The coincident datasets were then used to test and develop MLA to calibrate PA data. We tested several algorithms, and based on their performance, random forest (RF) was selected as a candidate algorithm for further analysis. The MLA model used PA channels A and B average $PM_{2.5}$, T, and RH as inputs and generated calibrated $PM_{2.5}$ as an output at an hourly time interval.

The statistical parameters examined include mean bias, slope, correlation coefficient, and percentage falling within certain error bounds. We analyzed the performance of raw and MLA calibrated $PM_{2.5}$ from PA sensors against those obtained from FEM as a function of various inputs to the MLA model. We also analyzed the PA channel A and B differences, sensor to sensor variability, and the PA's capability to capture diurnal cycle and day-to-day changes. Error characterizations are performed for both hourly and daily time averaging. The following are the main conclusions from our study:

- PA channel A and B measurements should be used to quality control the data before using it for any scientific analysis. We flagged all the hourly data as lower quality if the difference between the two channels is larger than 5 $\mu g\ m^{-3}$ in Raleigh, whereas this threshold is set to 5% for India.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022


Atmospheric Measurement Techniques Discussions

- The comparison between raw PA $PM_{2.5}$ and FEM under different pollution loading and environmental conditions showed few similarities. The Delhi data shows a correlation of 0.88 under high $PM_{2.5}$ concentration ($PM_{2.5} > 50$ µg m^{-3}) whereas it is only 0.34 under clean conditions (Raleigh, $PM_{2.5} < 20$ µg m^{-3}). This analysis suggests the need for regional and aerosol loading dependent calibration models (or correction equations) for PA data.
- The ML calibration models were developed and validated using a 10-fold cross-validation approach separately for the two regions. The calibrated PA $PM_{2.5}$ shows an excellent correlation ($R > 0.9$) with mean bias < 1 µg m^{-3} and RMSE of about 25.0 and 1.0 µg m^{-3} for Delhi and Raleigh, respectively.
- The biases in raw PA $PM_{2.5}$ show strong dependency on temperature and RH in Delhi, whereas it is weak and random in Raleigh. The calibrated PA data does not show any dependency on any of the input parameters to MLA.
- The calibrated data follows both diurnal and day-to-day cycles with almost no bias with respect to FEM $PM_{2.5}$. The raw PA-II also follows along with these cycles but with biases.
- More than 90% of calibrated $PM_{2.5}$ data sets fall within 25% and 15% of FEM $PM_{2.5}$ for hourly and 24-hourly averaging periods, respectively. This suggests a high accuracy of PA data after careful calibrations are applied.
- The mean absolute percent bias in calibrated data was within 10% and 5% for hourly and daily estimates of PM2.5, respectively (Table S1). Given that the typical uncertainties between monitor variabilities can range around ±10% or larger for reference methods (Chow et al., 2008), we consider 5-10% error as being quite low for the category of low-cost sensors and indicates exceptional promise.
- Our sensor collocation was, however, limited to the winter period in both regions. Therefore, the application of our models to other seasons and larger field deployment needs to be evaluated and refined. Sensor degradation with time is another aspect that is not addressed in this study.
- The calibration method developed in this study is specifically designed to address the data quality need of our citizen science project. This is in line with the conclusion drawn in the review article (Giordano et al., 2022) that correction or calibration method or level of error tolerance depends on LCS data applications.

Low-cost sensors, especially as part of citizen science initiatives, can help in bringing communities into the air pollution discourse and drive behavioral change among citizens. Moreover, bias-corrected data generated from these sensors can complement the limited regulatory monitors and can improve the knowledge of the spatial distribution of $PM_{2.5}$ concentrations and population exposure.

8 Data Availability

The processed data used in this study will shortly be made available through the project website (https://aqcitizenscience.rti.org/#/home) after agencies internal approvals. The raw PurpleAir data used in the study are publicly available from purpleair.com.

9 Acknowledgement

This work is funded by the NASA Citizen Science for Earth Systems Program (CSESP) through cooperative agreement No. 80NSSC18M0101 to RTI. International. We also like to thank Dr. Andrea Clements of US EPA for helping with sensor collocation at the EPA site and for reviewing and providing comments on the draft manuscript. The authors also thank the students at IIT Delhi who helped with the sensor and wireless router setup and maintenance during the collocation period. Special thanks to Paul Lin, who helped in running trained ML models for data analysis. SD acknowledges financial support for the Institute Chair fellowship.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

10 References

Adam, M., Putaud, J.P., Martins dos Santos, S., Dell'Acqua, A., Gruening, C., 2012. Aerosol hygroscopicity at a regional background site (Ispra) in Northern Italy. Atmos. Chem. Phys. 12, 5703-5717.

Badura, M., Batog, P., Drzeniecka-Osiadacz, A., Modzel, P., 2019. Regression methods in the calibration of low-cost sensors for ambient particulate matter measurements. SN Applied Sciences 1, 622.

Barkjohn, K.K., Gantt, B., Clements, A.L., 2021. Development and application of a United States-wide correction for PM2.5 data collected with the PurpleAir sensor. Atmos. Meas. Tech. 14, 4617-4637.

Bi, J., Stowell, J., Seto, E.Y.W., English, P.B., Al-Hamdan, M.Z., Kinney, P.L., et al., 2020a. Contribution of low-cost sensor measurements to the prediction of PM2.5 levels: A case study in Imperial County, California, USA. Environ. Res. 180, 108810.

Bi, J., Wildani, A., Chang, H.H., Liu, Y., 2020b. Incorporating Low-Cost Sensor Measurements into High-Resolution PM2.5 Modeling at a Large Spatial Scale. Environ. Sci. Technol. 54, 2152-2162.

Cheng, B., Wang-Li, L., 2019. Spatial and Temporal Variations of PM2.5 in North Carolina. Aerosol and Air Quality Research 19, 698-710.

Chow, J.C., Doraiswamy, P., Watson, J.G., Antony-Chen, L.W., Ho, SSH, Sodeman, DA, 2008. Advances in integrated and continuous measurements for particle mass and chemical, composition. J. Air Waste Manage. Assoc. 58, 141-163.

Di Antonio, A., Popoola, O.A.M., Ouyang, B., Saffell, J., Jones, R.L., 2018. Developing a Relative Humidity Correction for Low-Cost Sensors Measuring Ambient Particulate Matter. Sensors 18, 2790.

Feenstra, B., Papapostolou, V., Hasheminassab, S., Zhang, H., Boghossian, B.D., Cocker, D., et al., 2019. Performance evaluation of twelve low-cost PM2.5 sensors at an ambient air monitoring site. Atmos. Environ. 216, 116946.

Gupta, P., Doraiswamy, P., Levy, R., Pikelnaya, O., Maibach, J., Feenstra, B., et al., 2018. Impact of California Fires on Local and Regional Air Quality: The Role of a Low-Cost Sensor Network and Satellite Observations. GeoHealth 2.

Hagan, D.H., Kroll, J.H., 2020a. Assessing the accuracy of low-cost optical particle sensors using a physics-based approach. Atmos. Meas. Tech. 13, 6343-6355.

Hagan, D.H., Kroll, J.H., 2020b. Assessing the accuracy of low-cost optical particle sensors using a physics-based approach. Atmos. Meas. Tech. Discuss., 6343-6355.

Kelly, K.E., Whitaker, J., Petty, A., Widmer, C., Dybwad, A., Sleeth, D., et al., 2017. Ambient and laboratory evaluation of a low-cost particulate matter sensor. Environ. Pollut. 221, 491-500.

Magi, B.I., Cupini, C., Francis, J., Green, M., Hauser, C., 2019. Evaluation of PM2.5 measured in an urban setting using a low-cost optical particle counter and a Federal Equivalent Method Beta Attenuation Monitor. Aerosol Sci. Technol., 1-14.

Martin, R.V., Brauer, M., van Donkelaar, A., Shaddick, G., Narain, U., Dey, S., 2019. No one knows which city has the highest concentration of fine particulate matter. Atmospheric Environment: X 3, 100040.

Masih, A., 2019. Machine learning algorithms in air quality modeling. Global Journal of Environmental Science and Management 5, 515-534.

Prakash, J., Choudhary, S., Raliya, R., Chadha, T.S., Fang, J., George, M.P., et al., 2021. Deployment of networked low-cost sensors and comparison to real-time stationary monitors in New Delhi. J. Air Waste Manage. Assoc. 71, 1347-1360.

PurpleAir, 2022. PurpleAir Data Map, Available at https://www.purpleair.com/map?opt=1/mAQI/a10/cC0#1/20/-30, Accessed March 10, 2022.

https://doi.org/10.5194/amt-2022-140
Preprint. Discussion started: 16 May 2022

Shiva Nagendra, S.M., Khare, M., 2019. Source Apportionment of Ambient Particulate Matter During Winter Season in Delhi, Report submitted to the Delhi Pollution Control Committee, Government of National Capital Region, New Delhi. , IIT Madras, Chennai and IIT Delhi, New Delhi, New Delhi, India, Available at https://www.dpcc.delhigovt.nic.in/uploads/news/edc30867ffd362e8500f2b1686b55eac.pdf

Si, M., Xiong, Y., Du, S., Du, K., 2020. Evaluation and calibration of a low-cost particle sensor in ambient conditions using machine-learning methods. Atmos. Meas. Tech. 13, 1693-1707.

South Coast AQMD, 2022. AQ-SPEC-Sensor Evaluations, Available at http://www.aqmd.gov/aq-spec/evaluations, Accessed March 10, 2022.

US E.PA, 2016. Quality Assurance Guidance Document 2.12: Monitoring $PM_{2.5}$ in Ambient Air Using Designated Reference or Class I Equivalent Methods, EPA-454/B-16-001, Research Triangle Park, NC, Available at https://www3.epa.gov/ttnamti1/files/ambient/pm25/qa/m212.pdf

van Donkelaar, A., Martin, R.V., Brauer, M., Boys, BL, 2015. Use of Satellite Observations for Long-Term Exposure Assessment of Global Concentrations of Fine Particulate Matter. Environ. Health Perspect. 123, 135-143.

van Donkelaar, A., Martin, R.V., Brauer, M., Hsu, N.C., Kahn, R.A., Levy, R.C., et al., 2016. Global Estimates of Fine Particulate Matter using a Combined Geophysical-Statistical Method with Information from Satellites, Models, and Monitors. Environ. Sci. Technol. 50, 3762-3772.

Wallace, L., Bi, J., Ott, W.R., Sarnat, J., Liu, Y., 2021. Calibration of low-cost PurpleAir outdoor monitors using an improved method of calculating PM2.5. Atmos. Environ. 256, 118432.

Wang, W.-C.V., Lung, S.-C.C., Liu, C.-H., 2020. Application of Machine Learning for the in-Field Correction of a PM2.5 Low-Cost Sensor Network. Sensors 20, 5002.

Wang, Y., Li, J., Jing, H., Zhang, Q., Jiang, J., Biswas, P., 2015. Laboratory Evaluation and Calibration of Three Low-Cost Particle Sensors for Particulate Matter Measurement. Aerosol Sci. Technol. 49, 1063-1077.

Zheng, T., Bergin, M.H., Johnson, K.K., Tripathi, S.N., Shirodkar, S., Landis, M.S., et al., 2018. Field evaluation of low-cost particulate matter sensors in high- and low-concentration environments. Atmos. Meas. Tech. 11, 4823-4846.

Zusman, M., Schumacher, C.S., Gassett, A.J., Spalt, E.W., Austin, E., Larson, T.V., et al., 2020. Calibration of low-cost particulate matter sensors: Model development for a multi-city epidemiological study. Environ. Int. 134, 105329.